The datacenter as a computer

In the last few decades, computing and storage have moved from PC-like clients to smaller, often mobile, devices, combined with large internet services. Traditional enterprises are also shifting to cloud computing.

This is happening because:

- The user experience improves: no configurations or backups are needed and the service can be accessed from anywhere at anytime
- SaaS allows for faster application development
- Fixing and improving software is easier in your own datacentre
- The hardware deployment is restricted to a few, well-tested, configurations
- Server-side computing allows for the faster introduction of new hardware devices, such as accelerators and many application services can run at a low cost per user

Furthermore, some workloads require so much computing capability that they are a more natural fit in datacentres, like search services, machine learning and deep learning.

The trend toward server-side computing and widespread internet services created a new class of computing systems: **Warehouse-scale Computers** (or *WSC*s).

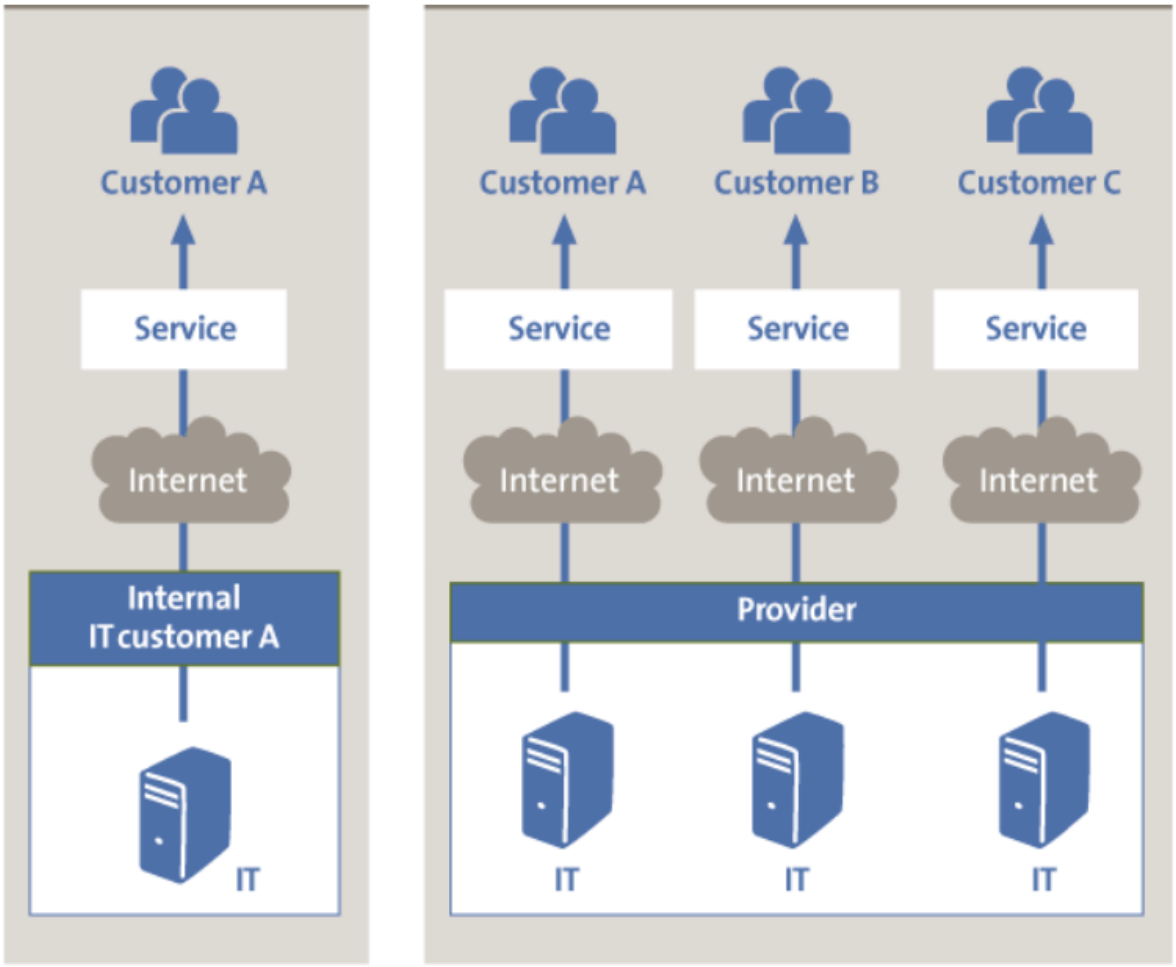
A program in a WSC:

- Is an internet service
- May consist of tens or more individual programs
- Such programs interact to implement complex end-user services, such as email, search, maps or ML.

Data centres are buildings where multiple servers and communication units are co-located because of their common environmental requirements and physical security needs.

Traditional data centres typically host a large number of relatively small or medium-sized applications. Each application runs on a dedicated hardware infrastructure

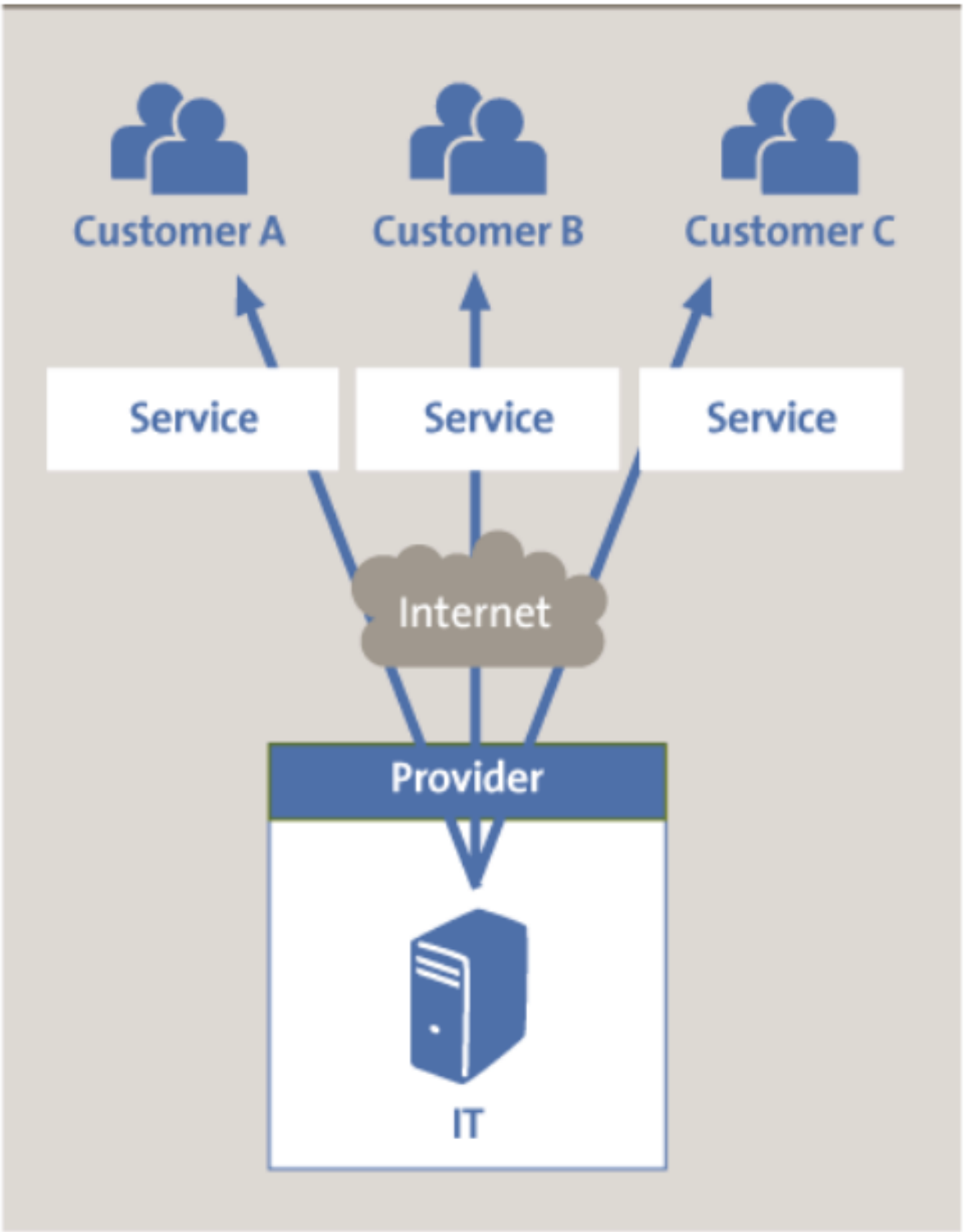
that is de-coupled and protected from other systems in the same facility, so applications tend not to communicate with one another.



WSCs belong to a single organisation, use a relatively homogeneous hardware and system software platform and share a common systems management layer.

WSCs run a smaller number of very large applications or internet services. The common resource management infrastructure allows significant deployment flexibility.

The requirements of homogeneity, single-organisation control and cost efficiency motivate designers to take new approaches in designing WSCs.



Initially designed for online, data-intensive web workloads, WSCs now power public cloud computing systems. Such public clouds do run many small applications, like a traditional data centre. All of these applications rely on VMs or containers and access large, common services for block or database storage, load balancing and so on.

The software that runs on WSCs executes on clusters of hundreds to thousands of individual servers, far beyond a single machine or a single rack.

Geographic areas and regions

In order to reduce user latency and improve throughput, companies usually build more than one data centre around the world. Often, each data centre is a replica of another one.

The world is divided into **Geographic areas** (*GA*s), defined by geo-political boundaries and determined mainly by data residency.

Each GA contains at least two **Computing regions** (*CR*s): customers see regions as the finer grain discretisation of the infrastructure. CRs usually have a latency-defined perimeter.

Availability zones (*AZ*s) are finer grain locations within a single computing region. They allow customers to run mission critical applications with high availability and fault tolerance to data centre failures.

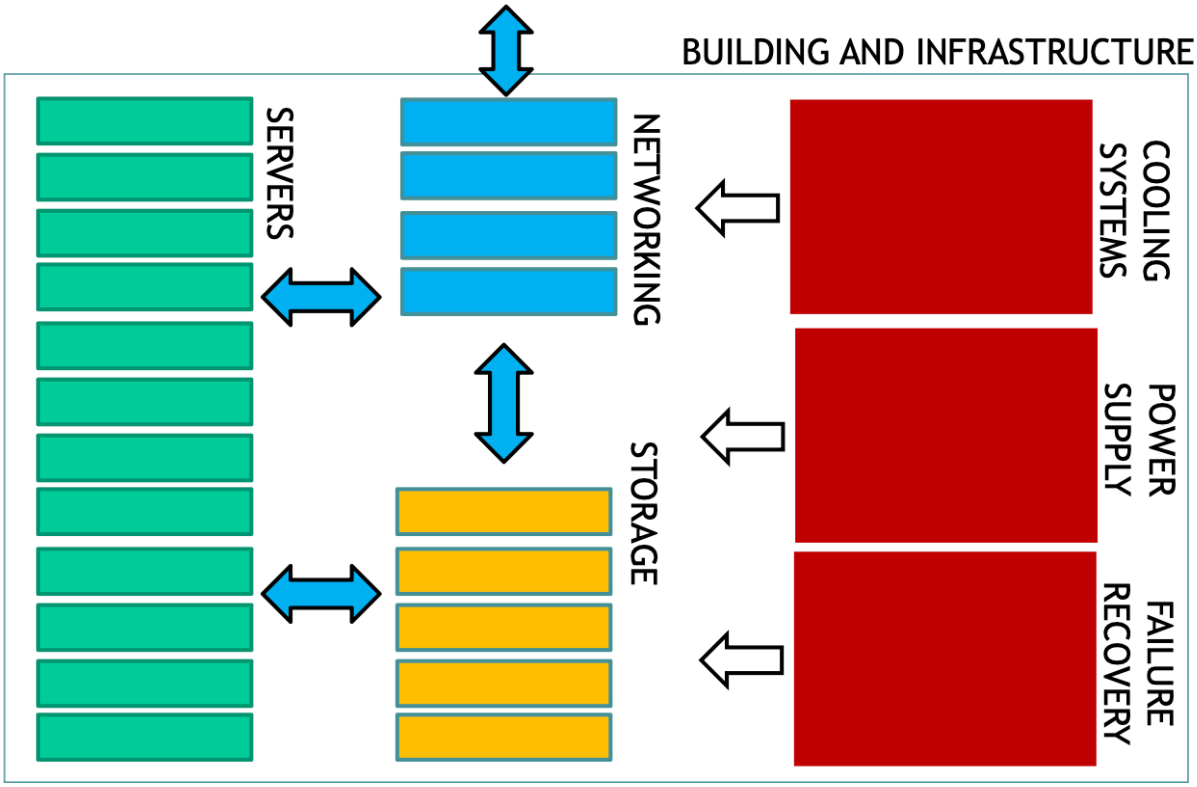
Application-level synchronous replication is possible among AZs, since they are much nearer than CRs or GAs. Having at least three AZs is necessary to have a quorum, in case one of the AZs goes offline.

WSCs and availability

Services provided through WSCs must guarantee high availability, typically aiming for at least 4-nines. Achieving such fault-free operation is difficult when a large collection of hardware and system software is involved.

WSC workloads must be designed to gracefully tolerate large numbers of component faults with little or no impact on service level performance and availability.

Architectural overview of a WSC



Servers

Servers are like ordinary PC, but with a form factor that allows to fit them into racks.

Servers may differ in:

- Number and type of CPUs
- Available RAM
- Locally attached disks
- Other special purpose devices, like GPUs

Storage

HDDs and SSDs are the building blocks of today's WSC storage systems.

These devices are connected to the data centre network and managed by sophisticated distributed systems.

Some examples are:

- Direct Attached Storage (DAS)
- Network Attached Storage (NAS)
- Storage Area Networks (SAN)

- RAID controllers

Networking

Communication equipment allows network interconnections among the devices.

Some networking equipment found in data centres is:

- Hubs
- Routers
- DNS servers
- DHCP servers
- Load balancers
- Technology switches
- Firewalls

Other components

WSCs have other important components related to power delivery, cooling and building infrastructure that also need to be considered.

Some interesting numbers:

- Data centres can be as big as 110 football pitches
- Data centres can draw 150 MW of power or more
- Data centres usually have at least 4-nines availability